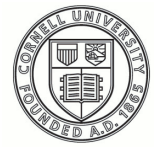


12-21-2015

Using Partially Synthetic Microdata to Protect Sensitive Cells in Business Statistics

Javier Miranda
Center for Economic Studies, US Census Bureau, javier.miranda@census.gov

Lars Vilhuber
Cornell University, lv39@cornell.edu

Using Partially Synthetic Microdata to Protect Sensitive Cells in Business Statistics

Abstract

We describe and analyze a method that blends records from both observed and synthetic microdata into public-use tabulations on establishment statistics. The resulting tables use synthetic data only in potentially sensitive cells. We describe different algorithms, and present preliminary results when applied to the Census Bureau's Business Dynamics Statistics and Synthetic Longitudinal Business Database, highlighting accuracy and protection afforded by the method when compared to existing public-use tabulations (with suppressions).

Keywords

synthetic data, statistical disclosure limitation, time-series, local labor markets, gross job flows, confidentiality protection

Comments

Presented at World Statistical Congress 2015 and Joint Statistical Meetings 2015.Vilhuber acknowledges support through NSF Grants SES-1042181 and BCS-0941226. All authors were affiliated with the U.S. Census Bureau, Center for Economic Studies, when originally contributing to the contents of this paper. This document reports the results of research and analysis undertaken by U.S. Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This document is released to inform interested parties of ongoing research and to encourage discussion of work in progress. All results have been reviewed to ensure that no confidential information is disclosed. The views expressed herein are attributable only to the authors and do not represent the views of the U.S. Census Bureau. The data used in this paper is restricted-access, and can be accessed either through the Federal Statistical Research Data Centers (LBD) or through the Synthetic Data Server at Cornell University (Synthetic LBD). Data and code used for the final version of this paper will be archived at the U.S. Census Bureau and made available upon request.

Using partially synthetic microdata to protect sensitive cells in business statistics

Javier Miranda and Lars Vilhuber *

Abstract. We describe and analyze a method that blends records from both observed and synthetic microdata into public-use tabulations on establishment statistics. The resulting tables use synthetic data only in potentially sensitive cells. We describe different algorithms, and present preliminary results when applied to the Census Bureau's Business Dynamics Statistics and Synthetic Longitudinal Business Database, highlighting accuracy and protection afforded by the method when compared to existing public-use tabulations (with suppressions).

Keywords: synthetic data, statistical disclosure limitation, time-series, local labor markets, gross job flows, confidentiality protection

* **Miranda:** U.S. Bureau of the Census, Washington, DC, USA `javier.miranda@census.gov`. **Vilhuber:** (Corresponding Author) Cornell University, Ithaca, NY, USA, `lars.vilhuber@cornell.edu`

1 Introduction

Statistics based on detailed business data are increasingly relied upon to make informed decisions by firms and governments. Novel statistics, for instance on business startups and firm dynamics [1], are valuable additions to the toolbox of evidence-based policy and business decisions. At the same time, the sparsity and skewness of business data makes disclosure avoidance a challenge. Early County Business Patterns (CBP) statistics (before the advent of noise infusion as a disclosure avoidance measure) had between 10 and 40 percent of values suppressed.[1]

In recent years, the use of fully or partially synthetic data has allowed the publication of increasingly detailed statistics. Going back to the seminal contributions of [2] and [3], the release of statistics based on partially synthetic data in the U.S. Census Bureau's LEHD Origin-Destination Employment Statistics (LODES) [4] and American Community Survey (ACS) [5–7] strikes a new balance between detailed statistics and appropriate disclosure avoidance. Other cases of using synthetic data for the purpose of tabulation exist [8, 9]. Furthermore, partially synthetic microdata [10–12] has been released to end-users as a access and analysis mechanism [13].

In this paper, we explore the use of tabulations based on partially synthetic data as a disclosure avoidance mechanism for certain at-risk tabulation cells. This is similar in spirit to the originally proposed uses of synthetic data [3, 2], and follows similar uses of partially synthetic data in the ACS [6, 7]. Our approach differs in that we address longitudinal consistency of the data explicitly, an important feature of the statistics underlying our paper.

To illustrate and implement the proposed mechanism, we use the Business Dynamics Statistics (BDS). The BDS were first released in 2008, providing novel statistics on business startups on a comprehensive basis for the U.S. economy [1]. They have been used in a number of recent publications, addressing questions of job creation and destruction, establishment births and deaths, and firm startups and shutdowns [14–17]. The BDS are sourced from confidential microdata in the Longitudinal Business Database (LBD). It provides measures of business openings and closings, and job creation and destruction, by a variety of cross-classifications (firm and establishment age and size, industrial sector, and geography). Since the first release, additional cross-tabulations have been added each year:

[1] Example taken from 2004 CBP, national by NAICS tabulations, across all size and NAICS cells.

initially provided only based on firm charateristics, tabulations based on establishment characteristics were later added, as were additional geography cross-tabulations (Metropolitan Statistical Area, and Metro/Non-Metro). Sensitive data are currently protected through suppression. However, as additional tabulations are being developed, at ever more detailed geographic levels, the number of suppressions increases dramatically.[2]

We leverage the existence of a sophisticated partially synthetic data file the Synthetic LBD [18, 11], henceforth SynLBD – in combination with the techniques first expressed in [19] and [20] to replace sensitive cells with tabulations based on synthetic data. A previous paper [21] described early results from the implementation of the simplest algorithm described here. In this version, we refine those algorithms, and present new results. We start by describing the extent of suppressions in the BDS, then lay out the algorithm to combine synthetic and confidential data for the purposes of tabulation. Preliminary results are discussed, and an outlook given on the next steps necessary to achieve a robust public-use tabulation.

2 Current Protection Methods

BDS processing uses primary and secondary suppressions, derived from a P percent rule, as disclosure avoidance mechanism. All cells of a potential publication table are analyzed to make sure no identifying information about a particular business, household, or individual is released to the public. In the case of the BDS, cells where the top 2 firms account for more than P percent of the total value of the cell are flagged for suppression. The precise P value is not disclosed to minimize the possibility of reidentification by potential attackers. Secondary suppressions are identified so as to minimize the amount of information loss in a given table row or column. To this end, the search algorithm looks for candidate cells that contain the least amount of employment, and suppresses their content. Protecting these secondary cells might require a third round of supressions given the presence of column totals in the tables. Once the tables are analyzed and the necessary cells suppressed, each table row that contains a suppression is flagged, and the modified table released to the public.[3] A necessary feature of this disclosure mechanism is that a large number of secondary suppressions are necessitated by the need to

[2] The next set of expansions include plans to provide additional industry and geography detail.

[3] Note that in some data release formats (SAS) individual suppressed cells are not separately flagged, only the row that contains at least one suppressed cell.

protect the cell that is the primary disclosing cell. The public-use data, of course, doesn't allow the identification of which suppressions are primary or secondary suppressions.

Table 1 describes the extent to which suppressions occur in the published establishment-level BDS [22] (Table 6 in the appendix also describes the similar pattern in firm-level statistics). The number of cells in each table is indicated, as are the percent of cells with suppression of some variable (`d_flag=1`), and the percent of cells where "Job Creation by Entrants" or "Job Creation by Continuers" is suppressed. Other variables, also present on the establishment-level BDS, are never suppressed.

Clearly, while the usefulness of the data to users would seem to increase for more detailed cross-tabulations, that same detail, under current disclosure avoidance rules, leads to increased suppression, and thus less effective data utility. Suppression is worse for some variables than for others. Establishment and firm counts are never suppressed following County Business Patterns and Disclosure Review Board rules. By contrast job creation and destruction, and establishment birth and deaths may be suppressed.

3 Alternative Protection Methods

In this section, we describe a protection system which uses tabulations from synthetic data, in a variety of implementations, to compensate for the suppressions generated by the current protection system. They should be considered extensions or complements to the current protection methods, since we describe and implement them within the constraints of the current system. In particular, our definition of sensitive cells is driven entirely by the current protection system. For comparison purposes, we also (partially) implement an alternate protection system, multiplicative noise infusion.

3.1 Synthetic Data Tabulations

The Synthetic LBD (SynLBD) [18] is a synthetic dataset on establishments with proven analytic validity along several critical dimensions [11]. Additional improvements are currently being developed [23, 24]. A growing number of researchers have used the SynLBD, and their continued use contributes to the improvement of the SynLBD.

The use of the SynLBD for the purposes outlined in this paper is particularly appealing, because its analytic validity has been independently established, while maintaining a high level of data privacy. Based

on the variables already available on the released SynLBD, tabulations that use the SynLBD as an input presumably require no additional disclosure avoidance review. Only tabulations involving state and sub-state geography should require additional review since geographic variables were removed from the disclosure request that approved the release to the public of the SynLBD.[4]

The available SynLBD is released as a single implicate, and by design, may distort an analysis by a potentially large an amount. The use of additional implicates for the purposes of BDS table creation may be desirable and will be assessed in later work.

In this paper, we propose and evaluate several algorithms that complement the existing BDS disclosure avoidance methodology (primary and secondary suppression, PSS). In all cases, we allow the PSS methodology to determine which cells are sensitive. Once identified, sensitive cells as well as some additional cells are modified using tabulations based on synthetic establishments.

The first algorithm, which we will call the "drop-in algorithm", simply replaces a cell that has been suppressed with its synthetic-data equivalent, i.e., the equivalent table cell from a tabulation based on the SynLBD alone. The second algorithm, called "forward-longitudinal algorithm", is slightly more complicated. At any point in time t, if a (expanded) suppression algorithm identifies a cell that *would* be suppressed under PSS, all establishments that contribute to that cell in time period t are replaced by synthetic establishments that match on certain characteristics Z in periods $t-p$ through t, for t and the next n periods. Synthetic and observed values are then tabulated to create the release statistics. To smooth the phasing out of the synthetic establishments, we define weights $w(n)$ that decline monotonically from unity to zero for synthetic establishments, and increase correspondingly for real establishments. If Z describes only the margin characteristics for the table in question (denoted by k below), and not any additional characteristics, and for $p = n = 0$, the algorithm is similar to the "drop-in" algorithm, but creates consistent higher-level tables automatically. On the other hand, the "forward-longitudinal algorithm" cannot be done post-publication without the re-release of historical tabulations.

In this paper, we will restrict ourselves to $p = 0$ and $n = \{0, 4\}$ in order to assess the time-consistency of the proposed algorithms for a single implicate. We have previously assessed the impact of Algorithm 1

[4] The Census Disclosure Review Board has not pronounced itself on the disclosure avoidance methodology proposed here as of December 2015.

(defined below) [21]. Assessing the impact of using multiple implicates as well as identifying acceptable values of Z, p, and n is deferred to future work.

3.2 Definitions

The variable of interest is establishment employment e_{jt}, with establishments indexed by j and years indexed by t. All other variables (job creation and destruction from establishment entry, exit, expansion and contraction) are derived from that basis. For instance, an establishment is "born" at time t if employment is positive for the first time in t:

$$birth_{jt} = \begin{cases} 1 \text{ if } \ e_{jt} > 0 \ \text{ and } \ e_{jt-s} = 0 \ \ \forall s \geq 1 \\ 0 \text{ otherwise} \end{cases} \tag{1}$$

We will denote aggregations using capital letters, so (national) employment is denoted as

$$E_{.t} = \sum_{j=1}^{J} e_{jt} \tag{2}$$

and (national) births are

$$Birth_{.t} = \sum_{j=1}^{J} birth_{jt}. \tag{3}$$

An establishment j has a vector of time-invariant and time-varying characteristics $k_t(j)$, such as industry and geographic location (time-invariant), but also derived characteristics, such as establishment or firm age and size. In a slight abuse of notation, $j \in K'_t$ describes the set of firms at time t such that $k_t(j) = k'$. Generically,

$$X_{k't} = \sum_{j \in K'_t} x_{jt} \tag{4}$$

describes the different aggregations across establishments having characteristics k' at time t, for instance aggregations by establishment age or metropolitan areas, referred to as "confidential BDS" ($\mathbf{BDS}^{conf}$).

For any establishment j, the synthesized version of variable x_{jt} (from a single implicate) is denoted $\tilde{x}_j t$. The vector $\tilde{k}_t(j)$ describes the set of characteristics when using the synthetic dataset, which will generally differ from $k_t(j)$ because time-varying derived characteristics such as age

and size will differ (at this time, neither industry nor geography are synthesized). We designate the set of establishments j with synthetic characteristics $\tilde{k}_t(j)$ as $\tilde{K}'_t$, and will refer to them as "synthetic establishments." Aggregations across synthetic establishments are

$$\tilde{X}_{k't} = \sum_{j \in \tilde{K}'_t} \tilde{x}_{jt} \tag{5}$$

and will be referred to as "synthetic BDS" ($\mathbf{BDS}^{(s)}$).

Finally, suppression rules for (aggregate) variable X are captured by I_t^X, such that the releasable variable X^o under the current regime (PSS) can be described by

$$X^o_{k't} = \begin{cases} X_{k't} \text{ if } I^X_{kt} = 1 \\ \text{missing otherwise} \end{cases} \tag{6}$$

For later reference, we denote the tabulations created as per (6) as $\mathbf{BDS}^{(0)}$.

3.3 Algorithm 1: Drop-in

We can now express the "drop-in" algorithm, leading to the released variable $X^{(i)}$, as:

---

if $I_t^X = 0$ **then**
  $X^{(i)}_{k't} = \tilde{X}_{k't}$
else
  $X^{(i)}_{k't} = X_{k't}$
end if

Thus, simply computing a "SynBDS", based on the SynLBD, in parallel to the computation of the BDS, based on the confidential LBD, and replacing suppressed cells with their fully synthetic counterparts, yields a dataset without missing cells. Note that we have assumed the existence of only one synthetic implicate; the use of multiple synthetic implicates would replace the second component of Algorithm 1 with $X^{(i)}_{k't} = \frac{1}{\ell} \sum_{l=1}^{\ell} \tilde{X}_{k'tl}$, the average across ℓ implicates. In general, increasing the number of implicates will improve the analytic validity, but reduce the protection provided by the synthesis process.

Because no time-consistency is imposed, this method can lead to seam biases or higher intertemporal variance. Furthermore, only interior cells

are adjusted, but no margins are corrected, likely leading to discrepancies in the global table structure. Raking would solve that issue, but is not explored here.

In order to smooth the time-series generated by this process, and to provide a comparison to the microdata-based smoothing outlined later in this section, we generalize the above algorithm to combine not just synthetic tabulations in periods with suppression, but also in later periods. Thus, in periods that follow a period with $I_t^X = 1$, we average synthetic tabulations with non-suppressed tabulations, for up to n periods:

Algorithm 1: Weighted Drop-in

$s^* = min_{s \in [0,n]}$ s.t. $I_{t-s}^X = 0$
if $n > 0$ and $\exists s^*$ **then**
  $X_{k't}^{(i)} = \frac{s^*}{n} X_{k't} + \left(1 - \frac{s^*}{n}\right) \tilde{X}_{k't}$
else if $n = 0$ and $I_t^X = 0$ **then**
  $X_{k't}^{(i)} = \tilde{X}_{k't}$
else
  $X_{k't}^{(i)} = X_{k't}$
end if

For $n = 0$ this reduces to the prior expression. For later reference, we denote the tabulations created by Algorithm 1 as $\mathbf{BDS}^{(i)}$ in its general form, and as $\mathbf{BDS}^{(in)}$ when $n = 0$.

3.4 Algorithm 2: Forward-longitudinal

In part to address the possible time-inconsistencies we propose an alternative algorithm. In order to minimize future seam issues, we downweight or remove establishments (or firms) that contribute to sensitive cells of tabulations with characteristics $k't$, for t and the next $n-1$ periods. These establishments are (partially) replaced by synthetic establishments that match on characteristics $k't$, and we simply replace the observed values in the database x_{js} with the synthetic values $\tilde{x}_{js}$ (for all variables), for $s = t, \ldots, t+n$. For convenience, denote by $J_{k't}^-$ the set of establishments that are to be excluded from tabulations at time t, and $J_{k't}^+$ the set of synthetic establishments that are added to the tabulations as replacements. We construct $J_{k't}^-$ by first adding establishment identifiers that meet the suppression conditions I_{kt}^X at time t. In addition, we assign establishments to $J_{k's}^-$ for the n periods after a cell stops being sensitive as well. Formally, we add those same establishments to "future" I_{ks}^X, for $s \in [t+1, t+n]$ if $n > 0$. Thus, at any point in time t, the set $J_{k't}^-$ contains establishments

that met suppression conditions now and in the past, i.e., in $[t-n,t]$. In order to "smooth" the tabulated data, we specify a per-establishment weight $w_{js} \in [0,1]$, applied to the observed data, that increases from 0 in t to 1 in $t+n$, and a per-establishment weight $\tilde{w}_{js}$, applied to the synthetic data, that decreases from 1 in t to 0 in $t+n$, thus "blending in" the real establishments, and "blending out" the synthetic establishments. Setting $w_{js} = 0, s \in [t, t+n-1]$ and $\tilde{w}_{js} = 1, s \in [t, t+n-1]$ effectively removes the real establishments from the tabulation, being completely replaced by the synthetic establishments. In its simplest form, the algorithm can be expressed as

---

Algorithm 2: Forward-longitudinal

Compute: $X_{k't} = \sum_{j \in K'_t} x_{jt}$
Compute: I_t^X
if $I_t^X = 0$ **then**
  // Suppression condition met for cell k'
  Assign all $j \in K'_t$ to $J^-_{k's}$ for $t \le s \le t+n$
  Assign all $j \in \tilde{K}'_t$ to $J^+_{k't}$ for $t \le s \le t+n$
end if
Compute:

$$X^{(iiw)}_{k't} = \sum_{j \in \left\{K'_t \cap J^+_{k't}\right\}} \tilde{w}_{jt}\tilde{x}_{jt} + \sum_{j \in K'_t \wedge j \in J^-_{k't}} w_{jt}x_{jt} + \sum_{j \in K'_t \wedge j \notin J^-_{k't}} x_{jt}$$

---

where the first component is the (possibly down-weighted) sum of synthetic data, the second component is the (up-weighted) sum of observed establishments in periods after they are no longer part of sensitive cells, and the third component is sum of establishments that were not part of sensitive establishments in the past (or outside of the window $[t-n,t]$).

For $n = \infty$, J_t^- is an absorbing set, which seems undesirable. For $n = 0$, this is similar to, but not equal to Algorithm 1. Note that in contrast to Algorithm 1, all higher level tabulations are consistent, since the replacement occurs at the microdata level, not at the tabulation cell level.

Consider the case for period s for which $I_s^X = 1$ and $I_{s-1}^X = 0$, i.e., the suppression conditions no longer apply. By assignment in period $s-1$, some LBD establishments are still assigned to $J^-_{k's}$, and some synthetic establishments are still part of $J^+_{k't}$. However, new LBD establishments that are identified by k' are counted in $X^{(ii)}_{k't}$ by virtue of the second

sum. Equivalently, establishments (synthetic or real) that move out of k' (because they age or grow out of the category) naturally drop out of $X^{(ii)}_{k't}$. Note that because we condition on $J_{k'}$, synthetic establishments that naturally exit tabulation cell k' are not counted toward an alternative tabulation cell $k^\star$, unless that cell is *also* a candidate for suppression. For reference, we denote the tabulations created by Algorithm 2 as $\mathbf{BDS}^{(ii)}$.

3.5 Multiplicative Noise Infusion

Multiplicative noise infusion was originally proposed by [25]. Implementations include the Quarterly Workforce Indicators (QWI) [26] and the CBP. We apply multiplicative noise to employment counts and payroll measures (although our analysis in this paper only focuses on employment-based measures).

Multiplicative noise is drawn for each establishment j from a bilateral ramp distribution:

$$p\left(\delta_j\right) = \begin{cases} \dfrac{1+b-\delta}{\left(b-a\right)^2} & , \ \delta \in \ [1+a, 1+b] \\ \dfrac{\delta-(1-b)}{\left(b-a\right)^2} & , \ \delta \in [1-b, 1-a] \\ 0 & , \ \text{otherwise} \end{cases} \tag{7}$$

where $a = c/100$ and $b = d/100$ are constants chosen such that the true value is distorted by a minimum of c percent and a maximum of d percent. This produces a random noise factor centered around 1 with distortion of at least c and at most d percent. Figure 1 depicts such a distribution. The noise factor is drawn only once, and retained for all time periods after the initial assignment. For this exercise, we set $c = 10$ and $d = 25$ percent as plausible numbers, for illustration only. Note that these numbers are in general confidential, and we have no knowledge of the actual parameters used in QWI and CBP. Both QWI and CBP use slightly more complex noise infusion algorithms that takes into account the firm structure and table structure, and include suppression for the smallest cells where multiplicative noise provides insufficient protection. None of those additional features are implemented here. We denote the tabulations protected by noise infusion as $\mathbf{BDS}^{(n)}$.

4 Analysis

We used SynLBD [18] together with confidential BDS microdata (as of June 2015) for BDS tabulations by establishment age and size (`bds_e_agesz`), creating $\mathbf{BDS}^{(s)}$ and $\mathbf{BDS}^{conf}$, respectively. For the published data $\mathbf{BDS}^{(0)}$, we used data from the September 2014 release. We note that the BDS microdata is thus of more recent vintage, and contains some improvements in the underlying data. This leads to certain discrepancies in the results, as will be evidenced in the tables. Using Algorithm 1 and combining $\mathbf{BDS}^{(0)}$ and $\mathbf{BDS}^{(s)}$, we created $\mathbf{BDS}^{(i)}$ and $\mathbf{BDS}^{(in)}$. Using Algorithm 2 and combining the microdata underlying $\mathbf{BDS}^{(s)}$ and $\mathbf{BDS}^{conf}$, we created $\mathbf{BDS}^{(ii)}$ with $n = 4$, linear w_{js}, and $p = 0$. Further variation of the weights w_{js} and of n lead to $\mathbf{BDS}^{(ii)}(w = 0) = \mathbf{BDS}^{(iiw)}$and $\mathbf{BDS}^{(ii)}(n = 0) = \mathbf{BDS}^{(iin)}$, respectively. We create $\mathbf{BDS}^{(n)}$ with $c = 10$ and $d = 25$ percent the brackets of the noise distribution.

The analysis is restricted to 1977-1999 because SynLBD version 2.0 is only available through 2001, and we chose to avoid any issues at the boundaries of the data. As noted in Table 1, about 26% of all cells in publicly available $\mathbf{BDS}^{(0)}$ have some suppression. For this version of the paper, we analyzed two variables, "Job Creation by establishment births" (`job_creation_births`) and "Job Creation by continuing establishments" (`job_creation_continuers`). Other variables, such as the number of establishments (`estabs`) and "Employment" (`emp`), which are never suppressed, serve as a benchmark.

4.1 Extent of protection

Protection of the table relies in large part on the fact that the data replacing the suppressions is itself synthetic, and released (in the case of the examples in this paper) or (potentially) releasable (for tabulations with geography) to a broad audience [27]. No establishment's observed data is released in the SynLBD, and only the industry distribution of establishments is preserved exactly.[5] A detailed analysis, based in part on a comparison of the confidential and synthetic microdata is provided elsewhere [11]. Very few synthetic values are close to the corresponding confidential values, and [11] conclude that the synthetic microdata is not

[5] To be precise, the number of establishments that ever exist within each 3-digit Standard Industry Classification (SIC) throughout the timeframe of the synthesis is preserved exactly. At any given point of time, though, that number will diverge from the confidential number.

disclosive of the confidential microdata. It follows that tabulations of non-disclosive microdata are themselves not disclosive.

We do, however, note that one particular attribute present on the confidential microdata - geography - was not included on the released synthetic microdata. Several of the tabulations listed in Table 1 and 6 are cross-tabulated by geography. Two options thus arise: (i) to release a version of the SynLBD with protected geography (see ongoing work [23]), and then use that version for tabulations; (ii) to create a non-released version of the SynLBD that may not satisfy the criteria for release at the microdata level, but does allow for the computation of releasable tabulations. Neither of these options are explored in the present article.

4.2 Aggregate differences

The present version of the SynLBD, created in 2011, has some small aggregate differences with the released BDS tables. Our analysis will not take any particular measures to alleviate the bias. Figure 2 shows aggregated `denom` (average employment), job creation, and job creation by establishment births and continuing establishments, and Figure 3 presents the percentage difference between the released data and the synthetic data at the most aggregated level, for the same variables. We note that whereas total employment is only marginally lower in the synthetic data at any point in time, synthetic job creation is significantly higher. The synthetic data underestimates job creation by establishment births, and overestimates job creation by continuing establishments. These points were originally highlighted elsewhere [11], and are being addressed in the next iteration of the SynLBD.

For comparison, there are no such differences at the most aggregated level between releasable and noise-infused tabulations (not shown). Percentage differences are less than two-tenths of a percent in all cases, as expected.

4.3 Analytical validity

We turn to an assessment of analytical validity. In order to assess the analytical validity of each of the methods, we focus on simple time-series properties of the $X_{k't}$. In particular, we estimate a AR(2) process for each of time-series generated by $X_{k't}$, $X^{(0)}_{k't}$, $X^{(s)}_{k't}$, $X^{(i)}_{k't}$, $X^{(ii)}_{k't}$, $X^{(iiw)}_{k't}$, and $X^{(iin)}_{k't}$. We then assess, for each statistic X under each of the regimes, the number of feasible regressions for $X_{k't}$ (for some values of k, data points may be missing because out-of-scope in certain time periods), and what fraction

of the feasible regressions can be replicated under the alternate regimes. Table 2 presents these results for a number of variables. Conditional on a feasible estimation, we tabulate the fraction of ρ_1 estimates that are statistically significant at conventional levels (Table 2).

Two measures of utility are also computed. We compute *coverage* as the percentage of regressions where the true ρ_1 lies within the confidence band around the coefficient estimated from the comparison ρ_1^* for each of the tables generated by the different algorithms. Let (L^*, U^*) be the 95% confidence interval for ρ_1^*. *Coverage* is the percentage of AR(2) estimates for which the "true" $\rho_1 \in (L^*, U^*)$. We generalize this measure as suggested by [28], and compute the *interval overlap measure* J_k. Consider the overlap of confidence intervals (L, U) for ρ_1 (estimated from the confidential data) and (L^*, U^*) for ρ_1^*. Let $L^{over} = \max(L, L^*)$ and $U^{over} = \min(U, U^*)$. Then the average overlap in confidence intervals is

$$J_k^* = \frac{1}{2}\left[\frac{U^{over} - L^{over}}{U - L} + \frac{U^{over} - L^{over}}{U^* - L^*}\right]$$

We then average J_k^* over all estimated AR(2) regressions. Results are summarized in Tables 4 and 5.

We start by noting issues surrounding establishment births in both the synthetic and different releases of the observed data. Establishment births are particularly sensitive to identifier linkages, and regular improvements to the BDS microdata occur. On the other hand, it is one of the more difficult events to synthesize. This leads to discrepancies both between the synthetic and the confidential data, and between the public-use data released in September 2014 and the (preliminary) confidential tabulations from the June 2015 BDS microdata.

We further note that the first two rows of each table serve as a control, reporting results for **emp** and **estabs**, never show any differences, since there are no observed suppressions. Job creation, which also is never suppressed, does show some differences in Table 5 between $\mathbf{BDS}^{(0)}$ and $\mathbf{BDS}^{conf}$, presumably due to data revisions.

Tables 4 and 5 paint approximately the same picture. The synthetic data is sufficiently different to distort inferences in our application further away from the results obtained from the confidential data. While the published data, despite having suppressed cells, has an average J_k of 92.6%, filling in suppressions together with smoothing ($n = 4$) yields an average J_k of 77.5%. Clearly this is being driven by the increased use of statistically different synthetic data, since setting $n = 0$ (and thus using less synthetic data in the tabulations) yields a higher average J_k. The

results obtained through Algorithm 2 are qualitatively better, with very little variation across the parameter variations. Given the data vintage differences noted above, it is not reasonable to compare $J_k^{(0)}$ and $J_k^{(ii)}$ directly until consistent input data can be used.

5 Concluding remarks

In this paper, we have described several alternate mechanisms to substitute for suppressions in small-cell tabulations of business microdata, with the goal of improving analytic validity while maintaining a sufficiently high standard of disclosure limitation. Neither mechanism fundamentally changes the existing suppression methodology, rather, the mechanisms work to fill in the holes created by the suppression methodology. In particular, the first methodology (Algorithm 1) can be used ex-post, after initial publication of tabulations with cell suppressions.

Leveraging the availability of a high-quality synthetic dataset (the Synthetic LBD) with proven disclosure limitation efficiency and analytic validity [11], the first method is very simple, but may suffer from seam biases and time-inconsistency. The second method aims to improve on that by "blending in" real establishments after the need for suppressions has disappeared, which may slightly reduce analytic validity in time periods where the strict application of the suppression algorithms would no longer impose any constraints, but improving on the time-series properties of the released data.

For reference, we have also used a noise-infused version of the BDS, and performs similarly if not better.

The (preliminary) results do not bear out our hypothesis that the use of microdata for prolonged periods of time improves the analytic validity of the data. However, we refrain from definitive conclusions at this time, due to differences in the underlying microdata that contaminate the current results. Current improvements in the upcoming next release of both the existing BDS (expected in late 2015) and a new release of the Synthetic LBD will need to be incorporated for a more consistent analysis. Clearly, the success of our proposed methods depends heavily on the analytic validity of the underlying synthetic data being used.

Recent developments to improve the micro-level analytic validity of the SynLBD [24] should improve the analytic validity of the mechanisms proposed here as well. We also compare our proposed mechanisms to the actual published, but otherwise unmodified BDS. Comparing post-publication improvements to a table with suppressions [29] will inevitably

lead to an apparent reduction in the utility of this particular approach. Finally, the approach relies on continuous availability of synthetic microdata with analytical validity. Other approaches rely on fewer data points, and thus may be favored due to lower implementation costs.

Acknowledgments. Vilhuber acknowledges support through NSF Grants SES-1042181 and BCS-0941226. We thank Jorgen Harris for able research assistance. This project would not have been feasible without repeated input from Saki Kinney and Jerry Reiter, and their valuable work on the Synthetic LBD. *Disclaimer.* All authors were affiliated with the U.S. Census Bureau, Center for Economic Studies, when originally contributing to the contents of this paper. This document reports the results of research and analysis undertaken by U.S. Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This document is released to inform interested parties of ongoing research and to encourage discussion of work in progress. All results have been reviewed to ensure that no confidential information is disclosed. The views expressed herein are attributable only to the authors and do not represent the views of the U.S. Census Bureau. *Data access.* The data used in this paper is restricted-access, and can be accessed either through the Federal Statistical Research Data Centers (LBD) or through the Synthetic Data Server at Cornell University (Synthetic LBD). Data and code used for the final version of this paper will be archived at the U.S. Census Bureau and made available upon request.

References

1. Haltiwanger J, Jarmin R, Miranda J. Jobs Created from Business Startups in the United States [BDS Brief]; 2008. Available from: https://www.census.gov/ces/pdf/BDS_StatBrief1_Jobs_Created.pdf.
2. Rubin DB. Discussion of Statistical Disclosure Limitation. Journal of Official Statistics. 1993;9(2):461–468.
3. Little RJA. Statistical Analysis of Masked Data. Journal of Official Statistics. 1993;9(2):407–426.
4. Machanavajjhala A, Kifer D, Abowd JM, Gehrke J, Vilhuber L. Privacy: Theory meets practice on the map. International Conference on Data Engineering (ICDE). 2008;.
5. Rodríguez R. Synthetic Data Disclosure Control for American Community Survey Group Quarters [presentation]; 2007.
6. Hawala S. Producing partially synthetic data to avoid disclosure. In: Proceedings of the Joint Statistical Meetings. American Statistical Association; 2008. Available from: http://www.amstat.org/sections/srms/proceedings/y2008/Files/301018.pdf.
7. Zayatz L, Lucero J, Massell P, Ramanayake A. Disclosure Avoidance for Census 2010 and American Community Survey Five-year Tabular Data Products. In: Proceedings of the Joint Statistical Meetings. American Statistical Association; 2009. Available from: http://www.amstat.org/sections/srms/proceedings/y2010/Files/307156_57962.pdf.
8. Abowd JM, Gittings K, McKinney KL, Stephens BE, Vilhuber L, Woodcock S. Dynamically consistent noise infusion and partially synthetic data as confidentiality protection measures for related time-series. Federal Committee on Statistical Methodology; 2012. Available from: http://www.fcsm.gov/events/papers2012.html.
9. Sakshaug JW, Raghunathan TE. Synthetic Data For Small Area Estimation In The American Community Survey. Center for Economic Studies, U.S. Census Bureau; 2013. 13-19. Available from: http://ideas.repec.org/p/cen/wpaper/13-19.html.
10. Drechsler J. New data dissemination approaches in old Europe synthetic datasets for a German establishment survey. Journal of Applied Statistics. 2012;39(2):243–265. Available from: http://dx.doi.org/10.1080/02664763.2011.584523.
11. Kinney SK, Reiter JP, Reznek AP, Miranda J, Jarmin RS, Abowd JM. Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database. International Statistical Review. 2011 December;79(3):362–384. Available from: http://ideas.repec.org/a/bla/istatr/v79y2011i3p362-384.html.
12. Abowd JM, Stinson M, Benedetto G. Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project. U.S. Census Bureau; 2006. Available from: http://www2.vrdc.cornell.edu/news/?p=308.
13. Jarmin R, Louis TA, Miranda J. Expanding the Role of Synthetic Data at the U.S. Census Bureau. U.S. Census Bureau, Center for Economic Studies; 2014. 14-10.
14. Haltiwanger JC, Jarmin RS, Miranda J. Who Creates Jobs? Small vs. Large vs. Young. National Bureau of Economic Research; 2010. 16300. Available from: http://www.nber.org/papers/w16300.
15. Hurst E, Pugsley BW. What do Small Businesses Do? Brookings Papers on Economic Activity. 2011;43(2 (Fall)):73–142. Available from: http://ideas.repec.org/a/bin/bpeajo/v43y2011i2011-02p73-142.html.

16. Decker R, Haltiwanger J, Jarmin R, Miranda J. The Role of Entrepreneurship in US Job Creation and Economic Dynamism. Journal of Economic Perspectives. 2014;28(3):3–24. Available from: http://www.aeaweb.org/articles.php?doi=10.1257/jep.28.3.3.
17. Pugsley BW, Şahin A. Grown-up business cycles. FRB of New York; 2015. 707, revised Sept 2015.
18. U S Census Bureau. Synthetic LBD Beta Version 2.0. Washington,DC and Ithaca, NY, USA: U.S. Census Bureau and Cornell University, Synthetic Data Server [distributor]; 2011. Available from: http://www2.vrdc.cornell.edu/news/data/lbd-synthetic-data/.
19. Gittings RK. Essays in labor economics and synthetic data methods [Ph.D.]. Cornell University; 2009.
20. Drechsler J, Reiter JP. Sampling With Synthesis: A New Approach for Releasing Public Use Census Microdata. Journal of the American Statistical Association. 2010;105(492):1347–1357. Available from: http://ideas.repec.org/a/bes/jnlasa/v105i492y2010p1347-1357.html.
21. Miranda J, Vilhuber L. Using Partially Synthetic Data to Replace Suppression in the Business Dynamics Statistics: Early Results. In: Domingo-Ferrer J, editor. Privacy in Statistical Databases. vol. 8744 of Lecture Notes in Computer Science. Springer International Publishing; 2014. p. 232–242. Available from: http://dx.doi.org/10.1007/978-3-319-11257-2_18.
22. U S Census Bureau. Business Dynamics Statistics 2012 release. Washington,DC: U.S. Census Bureau [distributor]; 2014. Available from: http://www.census.gov/ces/dataproducts/bds/data.html.
23. Kinney SK, Reiter J. SynLBD: providing firm characteristics on synthetic establishment data. World Statistics Conference; 2013.
24. Kinney SK, Reiter J, Miranda J. Improving the Synthetic Longitudinal Business Database. U.S. Census Bureau, Center for Economic Studies; 2014. 14-12.
25. Evans T, Zayatz L, Slanta J. Using Noise for Disclosure Limitation of Establishment Tabular Data. Journal of Official Statistics. 1998 december;.
26. Abowd JM, Stephens BE, Vilhuber L, Andersson F, McKinney KL, Roemer M, et al. The LEHD Infrastructure Files and the Creation of the Quarterly Workforce Indicators. In: Dunne T, Jensen JB, Roberts MJ, editors. Timothy Dunne, J. Bradford Jensen, and Mark J. Roberts. University of Chicago Press; 2009. .
27. Abowd JM, Vilhuber L. Synthetic Data Server; 2010. Available from: http://www.vrdc.cornell.edu/sds/.
28. Karr AF, Kohnen CN, Oganian A, Reiter JP, Sanil AP. A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality. The American Statistician. 2006;60(3):1–9.
29. Holan SH, Toth D, Ferreira MAR, Karr AF. Bayesian Multiscale Multiple Imputation With Implications for Data Confidentiality. Journal of the American Statistical Association. 2010;105(490):564–577. Available from: http://dx.doi.org/10.1198/jasa.2009.ap08629.

Appendix

Ramp algorithm in SAS

```
%let seed=12345;
%let a=.1;
%let b=.25;

data test;
do i = 1 to 100000;
x=ranuni(&seed.);
     if 0   lt x le 0.5 then y=(1-&b) + (&b - &a)*(2*(.5-x))**.5;
else if 0.5 lt x le 1   then y=(1+&b) - (&b - &a)*(2*(x-.5))**.5;
output;
end;
run;

/* test the distribution */
/*
 * Turns on standard graphics;
 */
ods graphics on / imagename="SGPlot" width=3000px;
/*
 * Creates a (full-size page) PDF;
ods PDF  file="SGPlot.pdf" notoc dpi=600 ;
 * Creates a SVG file;
ods printer file="SGPlot.svg";
*/
options center nodate nonumber;

ods printer printer=png file="SGPlotHiDef.png" dpi=600;

proc sgplot data=test;
histogram y/ binstart=0.7 binwidth=0.001;
xaxis display=(nolabel) label='label'
   min=0.7 max=1&b. values=(0.75 0.9 1.1 1.25);
yaxis display=(nolabel) ;
run;
ods printer close;
```

Tables

Table 1. Suppressions in establishment-level BDS

Type	Number of cells	Suppressions (%)		
			Job creation	
		Any	by entrants	by continuers
Age	337	0.3	0.3	0.3
Age-Initial Size	3033	18.5	14.2	14.2
Age-SIC	3033	3	2.9	2.9
Age-State	19023	3.3	3.2	3.2
Age-Size	3033	26.9	16.1	16.1
All	36	0	0	0
Initial Size	324	0.3	0	0
Initial Size-SIC	2916	19.8	6.5	6.8
Initial Size-State	18357	26.8	11.2	11.6
SIC	324	0	0	0
State	1836	0	0	0
Size	324	0.3	0	0
Size-SIC	2915	28.1	11.6	12.3
Size-State	18358	31.7	14.5	15

Note: Cells are year x categories, where the number of categories varies by published table.

Table 2. Analytic validity: Feasibility of AR(2) regressions

Variable	Number feasible	Percent Infeasible							
	$X_{k't}$	$X^{(s)}_{k't}$	$X^{(0)}_{k't}$	$X^{(i)}_{k't}$	$X^{(in)}_{k't}$	$X^{(ii)}_{k't}$	$X^{(iiw)}_{k't}$	$X^{(iin)}_{k't}$	$X^{(n)}_{k't}$
emp	90	0	0	0	0	0	0	0	0
estabs	90	0	0	0	0	0	0	0	0
estabsentry	64	59.4	0	0	0	0	0	0	0
jobcreation	90	0	0	0	0	0	0	0	0
jobcreationbirths	90	25.6	18.9	13.3	13.3	1.1	2.2	1.1	0
jobcreationcontinuers	81	0	6.2	0	0	0	0	0	0

Table 3. Analytic validity: AR(2) regressions with significant parameters

Variable	Percent significant								
	ρ_1	$\rho_1^{(s)}$	$\rho_1^{(0)}$	$\rho_1^{(i)}$	$\rho_1^{(in)}$	$\rho_1^{(ii)}$	$\rho_1^{(iiw)}$	$\rho_1^{(iin)}$	$\rho_1^{(n)}$
emp	0.256	0.2	0.256	0.256	0.256	0.256	0.256	0.256	0.244
estabs	0.267	0.178	0.267	0.267	0.267	0.267	0.267	0.267	0.267
estabsentry	0.109	0	0.063	0.078	0.078	0.109	0.109	0.109	0.078
jobcreation	0.178	0.1	0.178	0.178	0.178	0.178	0.178	0.178	0.167
jobcreationbirths	0.078	0.015	0.068	0.09	0.115	0.067	0.08	0.067	0.078
jobcreationcontinuers	0.21	0.111	0.184	0.16	0.247	0.173	0.173	0.16	0.173

Table 4. Analytic validity: AR(2) regressions: Coverage

Variable	Coverage							
	$\rho_1^{(s)}$	$\rho_1^{(0)}$	$\rho_1^{(i)}$	$\rho_1^{(in)}$	$\rho_1^{(ii)}$	$\rho_1^{(iiw)}$	$\rho_1^{(iin)}$	$\rho_1^{(n)}$
emp	88.9	100	100	100	100	100	100	100
estabs	88.9	100	100	100	100	100	100	100
estabsentry	92.3	90.6	90.6	90.6	100	100	100	100
jobcreation	82.2	100	100	100	100	100	100	100
jobcreationbirths	89.6	91.8	91	89.7	97.8	97.7	98.9	100
jobcreationcontinuers	76.5	100	81.5	87.7	87.7	88.9	86.4	100

Table 5. Analytic validity: AR(2) regressions: Interval overlap

Variable	Interval overlap							
	$J_k^{(s)}$	$J_k^{(0)}$	$J_k^{(i)}$	$J_k^{(in)}$	$J_k^{(ii)}$	$J_k^{(iiw)}$	$J_k^{(iin)}$	$J_k^{(n)}$
emp	83.4	99.4	100	100	100	100	100	97.7
estabs	80.4	97.6	100	100	100	100	100	97.8
estabsentry	78.7	82.6	82.6	82.6	100	100	100	95.8
jobcreation	73.3	94.4	100	100	100	100	100	96
jobcreationbirths	72.9	80.9	81.5	79.9	91.9	91.9	91.8	94.5
jobcreationcontinuers	70.7	92.6	77.5	81.6	85.1	85.3	85	95.9

Table 6. Suppressions in firm-level BDS

Type	Level	No. of cells suppressed	Percent suppressed
all	f	35	0
metrononmetro	f	70	0
sic	f	315	0
age	f	325	0
agemetrononmetro	f	650	0
st	f	1785	0
agemsa	f	118950	0.3
szmsa	f	153688	1.4
agest	f	18360	1.8
agesic	f	2925	2.8
isz	f	420	9
iszmetrononmetro	f	840	9.8
sz	f	420	10.2
szmetrononmetro	f	840	11.1
iszst	f	23205	16.1
szst	f	23205	16.2
iszsic	f	3780	18.7
szsic	f	3780	19.9
ageisz	f	3874	24.2
agesz	f	3843	26.6
ageiszmetro	f	7647	29.1
ageszmetrononmetro	f	7575	30.8
ageiszsic	f	31500	41.3

Note: Cells are year x categories, where the number of categories varies by published table.

Figures

Fig. 1. Empirical distribution of noise

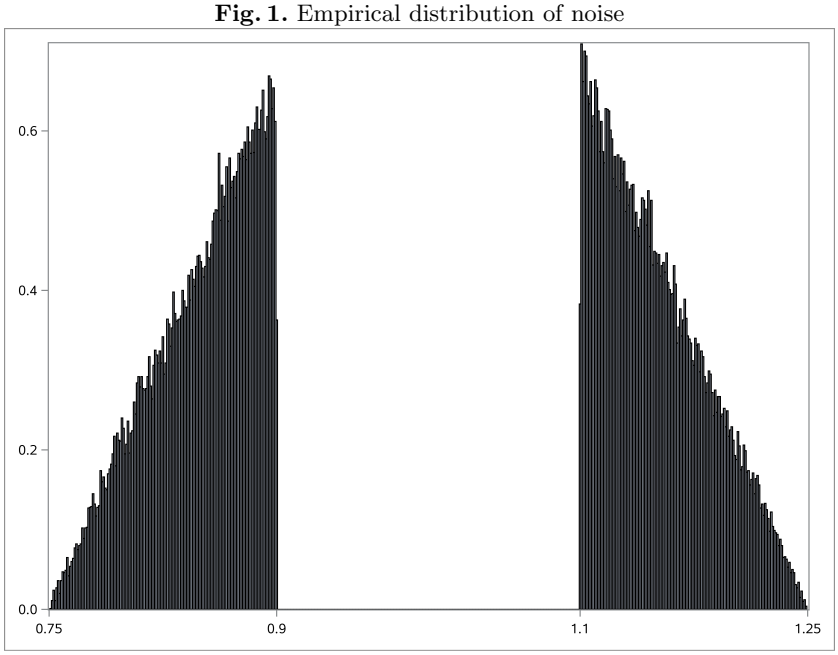


Fig. 2. Levels of released and synthetic data

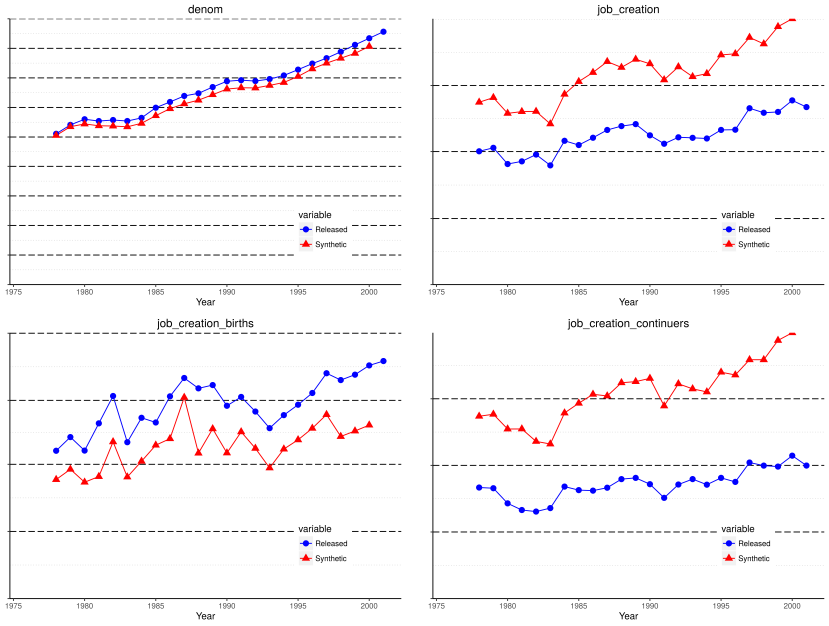
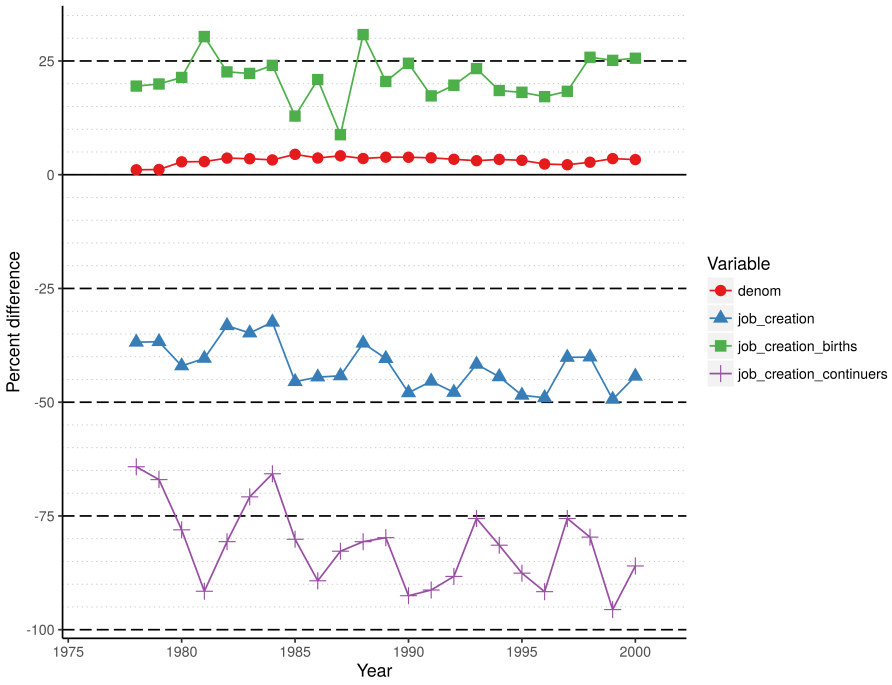


Fig. 3. Differences between released and synthetic data



This version: $Id: synbds-noise-synthetic-SJIAOS2015.tex 1784 2016-01-06 14:56:56Z lv39 $